DOCUMENT RESUME

ED 177 361 CE 023 066

AUTHOR Kershner, Keith
TITLE Research for Better Schools Career Education Analysis Service Guide. Education Is Going to Work.
INSTITUTION Research for Better Schools, Inc., Philadelphia, Pa.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE 78
CONTRACT NE-C-004-0011
NOTE 38p.; For related documents see CE 023 015, CE 023 064-067, CE 023 166, and CE 023 170-172
AVAILABLE FROM Research for Better Schools, Inc., 444 North Third Street, Philadelphia, PA 19123 ($3.00)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Career Education; *Evaluation; *Evaluation Methods; Guides; *Hypothesis Testing; Program Evaluation; *Statistical Analysis; *Statistics; *Tests of Significance

ABSTRACT

Designed for use with Research for Better Schools' evaluation system, this guide focuses upon basic concepts and statistical procedures applicable to specific research hypotheses emerging from evaluation questions. The first of three sections presents a discussion of descriptive statistics in terms of frequency, central tendency, variability, and applicability of descriptive statistics in evaluation. Focusing on inferential statistics, section 2 discusses samples and populations, hypotheses and probability, and applicability of inferential statistics in evaluation. The final section describes eight statistical tests: (1) one sample t-test, (2) independent samples t-test, (3) matched groups t-test, (4) univariate analysis of variance, (5) univariate analysis of covariance, (6) chi-square test, (8) product-moment correlation coefficient, and (8) rank-order correlation coefficient. (LRA)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED177361

CE023068

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

Ullik Roulk

RBS Publications

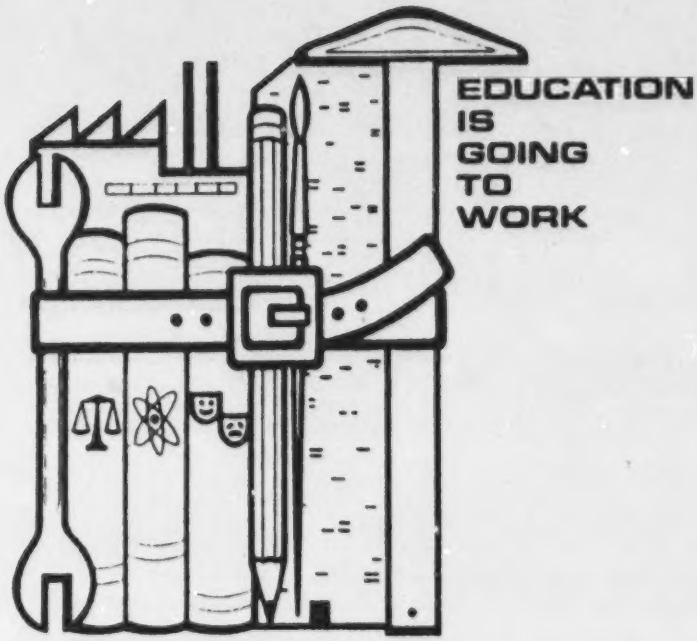
RESEARCH FOR BETTER SCHOOLS, Inc.
1700 MARKET STREET
PHILADELPHIA, PENNSYLVANIA 19103

RESEARCH FOR BETTER SCHOOLS
CAREER EDUCATION ANALYSIS SERVICE GUIDE

Keith Kershner

1978

2/a



RBS Career Education

ANALYSIS SERVICE GUIDE

RESEARCH FOR BETTER SCHOOLS, INC. (RBS), is a private, non-profit educational research laboratory located in Philadelphia, Pennsylvania. The ANALYSIS SERVICE GUIDE is part of a series of curriculum and procedural materials developed by the RBS CAREER EDUCATION PROGRAM (Michaelita B. Quinn, Director) for a pilot project in experience-based career education (EBCE). Additional materials in the evaluation series include:

EVALUATION PLANNING MANUAL

INSTRUMENT SERVICE GUIDE

PROGRAM MONITORING MANUAL

EVALUATION PACKAGE OVERVIEW

©Research for Better Schools, Inc. June 1977.

CONTENTS

	Page
INTRODUCTION	1
DESCRIPTIVE STATISTICS	3
1. Frequency	3
2. Central Tendency	6
3. Variability	9
4. Applicability of Descriptive Statistics in Evaluation	11
INFERENTIAL STATISTICS	13
1. Samples and Populations	13
2. Hypotheses and Probability	14
3. Applicability of Inferential Statistics in Evaluation	15
STATISTICAL TESTS	19
1. One Sample t-Test	19
2. Independent Samples t-Test	20
3. Matched Groups t-Test	21
4. Univariate Analysis of Variance	22
5. Univariate Analysis of Covariance	23
6. Chi Square Test	25
7. Product-Moment Correlation Coefficient	26
8. Rank-Order Correlation Coefficient	29
CONCLUDING NOTES	31
REFERENCES	32

INTRODUCTION

Evaluation has been a continuing activity in the development of RBS Career Education. The evaluation findings have provided a useful source of information in refining the program as well as offering evidence of program effectiveness to participants, sponsors, potential adopters and other members of the educational community.

Since RBS Career Education has become available for adoption by public school districts, a series of materials and services has been prepared to assist adopters in program planning, implemention and evaluation. This series includes:

Materials		Services
Evaluation Planning Manual	-	Evaluation Technical Assistance
Instrument Service Guide	-	Instrument and Scoring Service
Analysis Service Guide	-	Analysis Service
Program Monitoring Manual	-	Evaluation Technical Assistance

The materials are intended to assist in evaluation planning and design, while the services make available Research for Better Schools' evaluation systems and expertise to support implementation. These materials and services are described more completely in the RBS Evaluation Package Overview.

Any evaluation of program effects involves summarizing and analyzing one or more sets of data. Carefully planned analyses facilitate clear interpretation of the evaluation outcomes. The RBS Analysis Service Guide focuses upon basic concepts and statistical procedures which can be used to test the hypotheses established in the evaluation plan. It discusses underlying concepts in descriptive and inferential statistics which provide the framework for the analytical procedures and describes various statistical techniques in terms of contexts for use, output which may be generated and conclusions which can be drawn. Examples relating the statistical procedures to the evaluation of career education programs are presented wherever possible. A listing of basic references on the various analytic techniques and related concepts is given at the conclusion of the guide.

This guide can be of assistance in two ways. First, it can assist in evaluation planning by suggesting statistical procedures which might be appropriate for various program evaluations. Second, it can help staff members understand evaluation designs and findings by explaining some basic statistical techniques. The guide is not intended however, as a substitute for a statistical textbook and, in many cases, other assistance will be needed. Research for Better Schools can provide data processing services and evaluation technical assistance in planning, conducting and interpreting data analyses.

DESCRIPTIVE STATISTICS

Descriptive statistics summarize the distribution of scores among a particular group of subjects. They include measures of frequency, central tendency and variability.

1. FREQUENCY. Suppose an 80-item mathematics test has been administered to a class of 20 pupils. Since each item is scored as correct or incorrect, possible scores on the test range from zero to 80. The scores of the 20 pupils were as follows:

78	63	71	70
59	72	66	60
55	64	61	77
62	44	49	36
46	68	78	57

The evaluator may first construct an array of scores, displaying the distribution from highest to lowest:

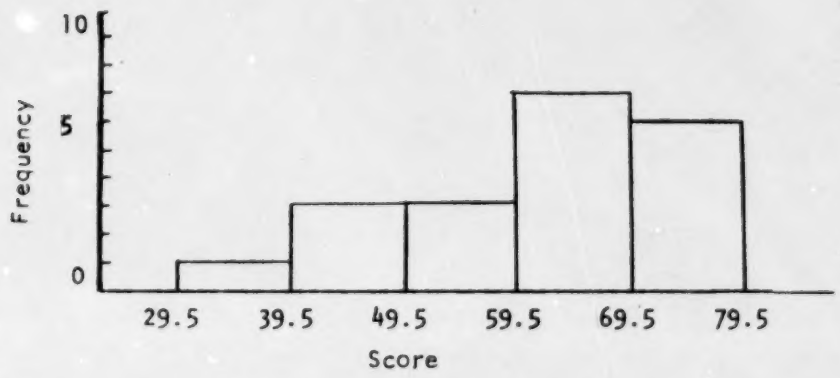
70's	60's	50's	40's	30's
78	68	59	49	36
78	66	57	46	
77	64	55	44	
72	63			
71	62			
70	61			
	60			

This rearrangement of scores tells the evaluator at least three things: the scores range from 78 to 36; the majority are clustered at the upper extreme (60's and 70's); and the "average" score appears to be in the 60's. With scores such as these, the second step would be to construct a frequency distribution, grouping the scores by class intervals.

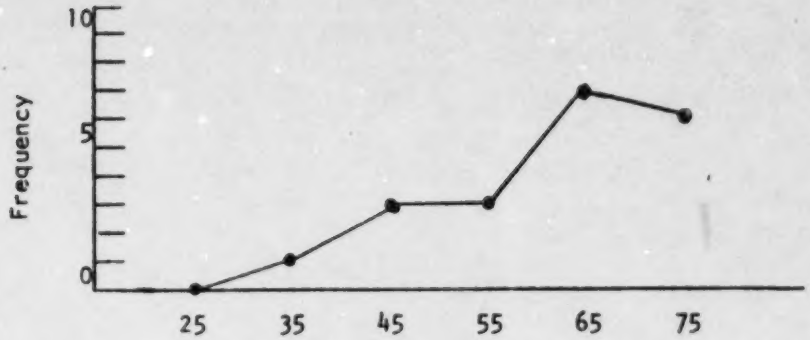
Class Interval	Frequency
69.5 - 79.5	6
59.5 - 69.5	7
49.5 59.5	3
39.5 - 49.5	3
29.5 - 39.5	1

Although individual score values are lost, the frequency distribution is a helpful summary when reporting large numbers of scores. The essential phase of construction is determining the appropriate class intervals. From the frequency distribution the evaluator can graph the distribution of scores on either a histogram or a frequency polygon.

HISTOGRAM



FREQUENCY POLYGON



Frequency tallies offer useful summaries of responses to interview and questionnaire items. For example:

Item: "The Career Education Program gave me a better idea of what I want to study in school."

	GRADES			
	9	10	11	TOTAL
Strongly Agree	11	15	6	32
Agree	27	14	17	58
Neutral	3	6	8	17
Disagree	5	8	4	17
Strongly Disagree	2	1	4	7
TOTAL	48	44	39	131

Since there may be unequal numbers of students in the different groups, it also may be helpful to present percentages of respondents falling into each category to obtain comparisons across groups.

Item: "Do you feel that all students should be allowed to smoke in designated areas of the building?"

	Students		Teachers		Parents	
	#	%	#	%	#	%
Yes	41	(82.0)	8	(40.0)	11	(44.0)
No	4	(8.0)	12	(60.0)	12	(48.0)
Unsure	5	(10.0)	0	(0.0)	2	(8.0)
TOTAL	50		20		25	

2. CENTRAL TENDENCY. There are three measures of central tendency in educational evaluation: the mean, the median and the mode.

The mean is the arithmetic average with which most people are familiar. When the "average weight" of the football team is given as 172 pounds, this probably refers to the mean. The statistical mean of this football team is computed as follows:

Name	Position	Weight
D'Orio	End	152
Anderson	End	181
Champaine	Tackle	188
Zipilli	Tackle	205
Freeman	Guard	172
Evans	Guard	161
Simkiss	Center Back	172
Smith	Back	152

Name	Position	Weight
Repplier	Back	162
Holloway	Back	166
Ferver	Back	182
		1893 Total

$$\text{Mean} = \frac{\text{Total of Scores}}{\text{Number of Scores}}$$

$$\frac{1893}{11} = 172.1$$

The median is the middle score of a distribution, or the point that divides the distribution in half. That is, half the scores fall above the median and half fall below. Computation of the median is simple with an odd number of scores because usually only one score occupies the center of the distribution:

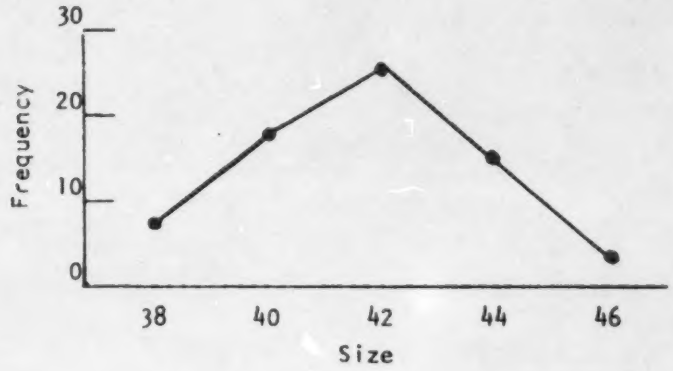
2, 3, 8, 9, 11, 14, 15

In the distribution above, nine is the median since it is the middle score. With an even number of scores, the median is calculated by taking the average of the two middle scores:

21, 29, 30, 32, 36, 42

In the above example the median would be 31. To calculate the median in more complicated situations, computational formulae are available in most elementary statistics books.

The mode is defined as the most frequently occuring score. For example, the modal suit-jacket size for American males used to be 40 regular, but is now 42 regular. For a group of 70 men we might find the following distribution:



The distribution above is called unimodal because there is one score that occurs most frequently. If there were two nonadjacent scores that occurred more often than other scores, the distribution would be called bimodal, or possessing two modes.

The mode can be useful when reporting free-response data from questionnaires or interviews, but otherwise it has minimal utility in educational evaluation. The mean, on the other hand, is useful in many situations, such as comparing test results of different groups or the same groups at different times. The median is best applied when there is a non-symmetric distribution of scores, which makes the mean a misleading indicator of central tendency.

For example, a single extreme score could distort the mean. In that case, the median would be more typical of the whole group:

Name	1975 Income
Baker	$ 5,500
Carpenter	7,900
Gregorio	8,300
Lemming	9,800
McNamara	$ 100,000

$$\text{Mean} = \frac{131,500}{5} = \$26,300$$

Median = $8,300

3. VARIABILITY. This term refers to the extent to which scores differ among themselves. Two distributions may have the same mean, but the scores may be distributed quite dissimilarly:

Group A	Group B
3, 3, 5, 6, 8 Mean = 5.0	1, 1, 3, 6, 14 Mean = 5.0

Group A's scores are clustered about the mean, while Group B's scores are spread out. Both the range and the standard deviation are commonly used to measure this spread, or variability.

The range is simply the arithmetic difference between the high score and the low score. The range of Group A above is 5 and the range of Group B is 13.

The standard deviation is a widely applied measure of variability that is generally used in conjunction with the mean. The standard deviation may be thought of as the average distance from the mean of scores in

a distribution. This statistic is computed as follows:

$$\text{Standard Deviation} = \sqrt{\frac{\text{Sum of squared differences from the mean}}{\text{Number of scores minus 1}}}$$

In the following example, the standard deviations of Group A and Group B have been computed. Although the means of both groups are identical, there is a difference in the way the individual scores are dispersed. Because Group B is more variable, it has a higher standard deviation (5.43, compared with 2.12).

Group A			Group B		
Score	Difference from Mean	Squared Difference from Mean	Score	Difference from Mean	Squared Difference from Mean
3	(3-5) = -2	4	1	(1-5) = -4	16
3	(3-5) = -2	4	1	(1-5) = -4	16
5	(5-5) = 0	0	3	(3-5) = -2	4
6	(6-5) = 1	1	6	(6-5) = 1	1
8	(8-5) = 3	9	14	(14-5) = 9	81
		SUM = 18			SUM = 118

$$\text{Standard Deviation} = \sqrt{\frac{18}{(5-1)}} = \sqrt{\frac{18}{4}} = \sqrt{4.5} = 2.12$$

$$\text{Standard Deviation} = \sqrt{\frac{118}{(5-1)}} = \sqrt{\frac{118}{4}} = \sqrt{29.5} = 5.43$$

The maximum value that a standard deviation can assume is approximately one-half the range. The minimum value is 0, meaning no deviation or no

difference among the scores (all scores are exactly the same). The standard deviation squared is known as variance.

The measurement of variability has several practical applications. Curriculum decisions can be based not only on the average performance of students, but on their range of performance as well. For example, a curriculum developed for students who perform at approximately the same level might be quite different from one developed for students who perform at disparate levels, even though the average performance in both situations is identical. In this case, variability is useful in program planning.

It may be a goal of a particular curriculum to decrease or increase the differences in achievement among students. In this case, measurements of variability would be used to determine program effectiveness.

Variability can also affect statistical relationships. For example, correlation coefficients between variables can be lowered by restricted standard deviations. A variable with a low standard deviation usually cannot correlate highly with anything.

4. APPLICABILITY OF DESCRIPTIVE STATISTICS IN EVALUATION. The first level of data analysis is always descriptive. Appropriate distributional statistics should be displayed for each subject group on all available measures. These data serve to depict group characteristics and suggest between-group differences which may need to be considered. Descriptive statistics answer some of the basic questions of interest to the evaluator:

- Is the posttest mean different from the pretest mean?
- Did the experimental group score higher or lower than the comparison group?
- Was the variability of scores especially high or low?
- Were the scores distributed differently in the different groups?
- Did certain groups of subjects do better than others on particular subtests or subscales?

From inspection of frequencies, means and standard deviations the evaluator can derive complete answers to some evaluation questions. For example, descriptive data can often answer questions about the extent of program implementation, participant opinions about the program, or program attendance levels.

In other cases, descriptive statistics can provide only preliminary answers. For example, the posttest means for two student groups involved in an educational program were as follows:

	Group A (Control)	Group B (Experimental)
Reading Comprehension	14.4	14.1
Arthmetic Concepts	16.1	18.2
Arthmetic Applications	9.1	16.0

The experimental group clearly scored higher on two of the three measures and slightly lower on the third. These descriptive results suggest that, if the pretest levels were equivalent, the posttests show a superiority

of the experimental group. However the educational significance of such a finding would have to be interpreted further, through the use of inferential statistics.

INFERENTIAL STATISTICS

Inferential statistics provide the means for testing research hypotheses. With descriptive statistics, the researcher is concerned solely with describing the scores (means, standard deviations, etc.) of a particular sample. In contrast, with inferential statistics the researcher is concerned with generalizing from sample characteristics to population characteristics.

1. SAMPLES AND POPULATIONS. A sample is defined as a particular group of individuals chosen as subjects for a research study. A population is defined as the general group from which the subjects are selected, that is, the group they represent. For example, the population of interest may be foreign-born females over 55 years of age, and from this population one randomly selects a sample of 100 women, If all members of a population were included in the sample, there would be no need for inferential statistics, since population values would already be known. Such is rarely the case, however, and the investigator is forced to make a statistical inference about the population.

Practical examples of statistical inference are well known. For example, the Neilsen ratings are not based on observations of the viewing habits of every household in the nation. Rather inferences about their viewing habits are based on a random sample of American households (the population of interest). Clearly, if the procedure for obtaining a

sample is inappropriate or flawed, the inferences drawn will be correspondingly limited, despite the use of sophisticated statistics.

2. HYPOTHESES AND PROBABILITY. The following example illustrates the test of a research hypothesis. Suppose one wanted to find out if a coin were a fair one; that is, if it were flipped, the chance of a head or tail appearing were equal. Suppose the experiment consisted of flipping the coin ten times. Hypothesizing that the coin is fair, five heads and five tails might be expected. If ten tosses yielded nine heads and only one tail, it might be concluded that the coin is unfair. If seven heads and three tails had been observed, it also might have been concluded that the coin is unfair, even though this latter results seems less decisive. Seven heads and three tails would be closer to the standard of fairness - five heads and five tails. One of the important functions of a statistical test is to determine whether observed values reliably support or refute hypotheses of interest.

In an experiment the reliability of support which can be lent to hypotheses by the obtained results is gauged by the probability of results occurring systematically rather than by chance. For the coin tossing experiment above, the number of heads one could observe in ten tosses of a coin is between zero and ten. Even if the coin were fair, it would still be possible to observe ten heads in ten tosses of a coin. However, it is a very rare event, likely to occur approximately once in a thousand repetitions of the experiment.

The researcher decides how rare an event must be for the hypothesis to be supported or rejected. In practice, a rare event is usually defined as one that would occur by chance either one in 10 times, one in 20 times or one in 100 times. These probabilities are referred to as levels of significance and are specified as the .10, .05 and .01 levels, respectively. Therefore, a result that is statistically significant at the .05 level means that it would occur by chance alone no more than five times in every 100 repetitions of the experiment.

The hypothesis in this example is that the number of heads equals the number of tails. It is this hypothesis that the researcher attempts to disprove or reject. If the hypothesis is rejected, then one concludes that there is a statistically significant difference between the number of heads and tails, and the coin is not fair. Alternative hypotheses are statements of an outcome or outcomes that include possibilities not covered by the initial hypothesis. In the coin example, an alternative hypothesis is that the coin is not fair and therefore the number of times It should come up heads is not equal to the number of tails.

3. APPLICABILITY OF INFERENTIAL STATISTICS IN EVALUATION. Suppose one were interested in whether a particular educational program has an effect on tenth grade students of all ability levels (the population of interest). Since it would be expensive, unwieldy and impractical to

include the entire population in the evaluation, a sample is randomly selected from the population and separated into two groups, experimental and comparison. At the end of the school term, both groups are tested to determine if the program effected any differences in student performance. If the program had no effect, no difference in posttest group means would be expected. An alternative hypothesis would be that there is a difference in favor of one group. Suppose that the mean of the experimental group was found to be 91.5 and that of the comparison group 81.2. The observed difference of 10.3 points must then be tested statistically to determine its reliability, or the probability that it occurred by chance rather than as a result of the program. If the difference in favor of the experimental group is found to be statistically significant at a predesignated level, then support is given to the hypothesis that the program is effective. If the difference is not significant, then no effectiveness can be inferred.

To apply any statistical test, the investigator must:

- Determine which test of significance is appropriate. (Eight frequently used tests of significance will be discussed in the next section of this guide).
- Select the appropriate level of significance, usually .01, .05 or .10.
- Determine number of degrees of freedom based on the number of cases (subjects) and groups of subjects being compared.
- Look up in the appropriate statistical table the value of the test statistic that must be exceeded in order for the obtained value to be significant. (These tables are found in most standard textbooks on inferential statistics).
- Calculate the test statistic by applying the appropriate formula or submitting the data to computer analysis.

In analyzing data, statisticians deal with a variety of situations, many of which are more complex than the examples presented above. In each situation, the assumptions made, the research question being investigated and the types of data being collected may differ. As examples of analyses are described in the next section, reference is made to some of the assumptions, hypotheses and data types that form the basis for the analysis.

STATISTICAL TESTS

This section describes eight commonly used statistical analyses. These descriptions are intended to perform an introductory function only. More information should be obtained before attempting to use or interpret the analytic techniques. The procedures covered include *t*-tests, analyses of variance and covariance, the chi-square test and correlation coefficients.

1. ONE SAMPLE t-TEST. The one sample, or single sample, *t*-test is a statistic for determining whether a group's mean score is significantly different from an established value. For example, assume that a class has taken a math test and that the class mean on the test was 68. Further, assume that a score of 70 is considered to be acceptable for students similar to those in the class. To determine whether or not class achievement is significantly below the acceptable level, a single sample *t*-test would be performed.

The *t* value that is calculated may be compared to critical values of *t* presented in the appropriate statistical table. The table entry that is of interest may be found by looking at the degrees of freedom (df), one less than the number of students (N-1), and the chosen level of significance (.01, .05 or .10). Under these levels in the table is listed the *t* value that must be exceeded for the result to be statistically significant. For instance, if the obtained value of *t* is larger than the tabled

value at the .05 level, it may be concluded with 95 percent confidence that the class score is significantly different from the established acceptable score. That is, a difference as large or larger than that which is obtained could have occurred by chance alone only five times out of 100.

2. INDEPENDENT SAMPLES t-TEST. In order to compare the performance of two independent samples, the independent sample t-test is used in a way that parallels to the one sample t-test. For two independent samples (two groups of comparable students), the question is whether or not the measured performance of one group is different from that of the other. As an example, the effectiveness of a reading program may be in question. Assume that all fifth graders in a school have been randomly assigned either to an experimental program or to the traditional reading class. After the treatment has occurred, the reading subtest of the Comprehensive Tests of Basic Skills is administered to determine whether the experimental reading program was any more or less effective than the control group treatment.

In this case the calculated t value is compared with the critical value given in the appropriate table under the selected level of significance and the appropriate degrees of freedom: N-2, or the total number of students minus two for the groups. If the absolute value of t exceeds the table entry, there is a statistically significant difference. If the absolute value of t is less than or equal to the tabled value, there is no significant difference between the two groups.

3. MATCHED GROUPS t-TEST. The matched groups t-test, sometimes called the correlated t-test, is designed to deal with two kinds of analysis: (1) the same group of students is observed on two different occasions by the same procedure; (2) two groups of students are compared, but each student is matched to a student in the other group on the basis of relevant variables such as age, ability, aptitude, and so on. Even though two sets of scores are available, they have been obtained from the same pupils or from two matched sets of pupils, and therefore a matched groups t-test must be used rather than the independent groups t-test.

An example of the first type of study in which the use of a matched groups t-test is appropriate would be as follows. As part of the evaluation of a career education program, a researcher administered the Career Maturity Inventory to 143 students in an experimental career education program both before the program began and after one year of operation. The researcher wants to determine to what extent students develop in career maturity over the course of a year.

To answer this the researcher can compare the mean score of the students after one year in the program with their mean score prior to the program experience. As noted above, the two sets of scores have been obtained from the same pupils and therefore a matched groups t-test must be used and not the independent samples t-test.

The second class of problems that the matched groups t-test is designed to handle involves the use of two different groups that have been matched with respect to variables the researcher wants to control.

For example, data might be collected regarding attitude toward a school policy. The researcher might be interested in whether students in two different schools differ in terms of their attitudes. One strategy that could be used is to obtain student samples from each school, administer and score the attitude scale, and then analyze the data using the matched groups t-test. The matching procedure serves as a control for the academic abilities of the sample students. That is, one might want to match students in the two groups on this variable, since academic ability might well be related to the attitude toward school variable. Then, if the results show different attitudes in the two schools, they cannot be a function of different academic ability in the two groups.

Following data collection, the value of t is calculated and compared to the table value. The degrees of freedom associated with the matched groups t-test is N-1, where N is the number of score pairs. The researcher chooses the appropriate significance level.

4. UNIVARIATE ANALYSIS OF VARIANCE. The independent samples t-test is designed to deal with data from two separate subject groups. The univariate analysis of variance is very similar to the independent samples t-test; the main distinction is that more than two groups are involved in analysis of variance procedures. Although there are more complicated analysis of variance techniques, only the "one-way" procedure will be discussed here. In this application, a researcher has either: (1) assigned a number of pupils to at least three different groups or (2) observed a number of independent groups. The number of students in each

group need not be the same, although they often are. The analysis of variance tests the significance of differences observed among the groups.

Suppose that as part of an evaluation of a career education program data on the Occupational Information subtest of the Career Maturity Inventory were collected for 195 pupils in grades 9-12. The mean scores for the four grades were 13.20, 14.44, 13.87 and 15.57. If a researcher were interested in testing the hypothesis that these four sets of scores are statistically different, the one-way analysis of variance would be appropriate. The analysis yields an F-ratio, which represents the measured differences between groups in relation to the variation within groups. The F-ratio is applied in a manner similar to the t value. Referencing the degrees of freedom for both subjects and groups, table values are compared with obtained F values. When the appropriate table value is exceeded by the obtained value, then results are concluded to be statistically significant.

5. UNIVARIATE ANALYSIS OF COVARIANCE. In analyzing educational program effects researchers expect that subject groups will be similar prior to their involvement in the program. This similarity is desirable so that one may have confidence that any differences observed among the groups after treatment will be a function of the program and not pre-existing differences in the groups. Occasionally, important differences do exist before the research study begins, and in these cases it is possible to compensate for them statistically by using the analysis of covariance.

The term covariance describes the tendency for two variables to increase or decrease together, e.g. height and weight. That is, tall people tend to weigh more than short people and vice versa.

The presence of group differences at the beginning of an evaluation study means that any differences found at the end of the study could have been caused by the initial differences rather than program effects. Thus, if the higher scoring students were prevalent in the experimental group, it would not be surprising for this group's mean score to be higher at the end of the study. The analysis of covariance may be applied to compensate for the initial differences that were found at the beginning of the study. It must be stressed that for an analysis of covariance to be valid the initial group differences must be due to unintentional sampling error, not systematic differences between the groups.

The covariance technique statistically equates the mean scores for the experimental and control groups obtained at the beginning of the experiment, and then allows for the calculation of adjusted posttest scores and an F-ratio based upon the adjusted scores. As with the analysis of variance, F values yielded by covariance procedures are checked in the appropriate table using degrees of freedom for subjects and groups. When more than two groups are involved in the analysis, additional procedures are needed to determine the statistical status of each group mean relative to the others.

6. CHI SQUARE TEST. Examining the number of subjects or responses that fall into various categories is often necessary in answering evaluation questions. Examples of this kind of study include the analysis of one school's racial distribution as compared to that of another school, or the analysis of teacher opinions about an important issue compared with parent opinions. The chi square test provides a way of analyzing the numbers of subjects, or responses observed in categories compared with those which could be expected to be found in the categories. The expected numbers are based on a direct proportional allocation. The observed numbers are those measured by the researcher. Where the observed values differ from expected values, the difference may be due to a treatment or variable of interest in the study.

In an example using only one subject group, assume that a school principle wants to determine if student absenteeism varies according to the day of the week. In this case, the observed frequency (value) would be the number of classes cut each day. The data may look like those presented below. The expected frequency (value) for each day would be the total number of observed cuts divided by the number of categories (days in the week). The chi square statistic will be zero in cases where observed and expected frequencies are the same. The larger the value of chi square, the more the observed and expected frequencies differ.

Number of Classes Cut

According to Day of the Week

	Monday	Tuesday	Wednesday	Thursday	Friday	Total
Observed Frequency	12	8	8	10	12	50
Expected Frequency	10	10	10	10	10	50

In this case the chi square value is calculated at 1.6 and there are four degrees of freedom (number of categories minus one). If this value is checked in the chi square table, it will be found to be not significant. That is, the hypothesis that there are no significant differences in the number of classes cut according to the day of the week is supported.

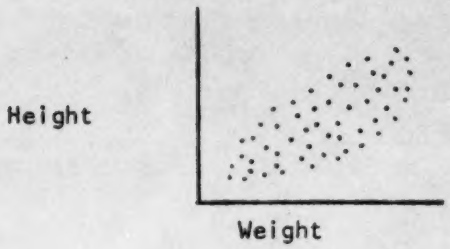
7. PRODUCT-MOMENT CORRELATION COEFFICIENT. Chi square deals with the extent of relationship between categorical variables. In order to examine the extent to which variables measurable on ratio or interval scales are related, the product-moment correlation coefficient is calculated. Ratio and interval scales contain definable points that have equal distances

between them. A ratio scale has a non-arbitrary zero point, whereas interval scales have an arbitrary zero point. Temperature measurement usually employs an interval scale and weight represents a ratio variable. Scores on standardized tests are usually interval data.

In performing correlations two measures are obtained from a group of subjects or from matched pairs of subjects. The obtained correlation coefficient gives an index of interrelationship that may range in value from -1.00 to +1.00. These extreme values represent a perfect linear relationship between two variables, while a zero correlation indicates that no relationship exists. Positive correlation coefficients are obtained when individuals who score high on one measure also obtain high scores on the second. Similarly, those with low scores on the first measure will tend to have low scores on the second measure. When individuals who get high scores on one variable get low scores on the second, and vice versa a negative correlation results. Some examples of correlations that may be of interest in program evaluation include those between attendance and achievement, ability and performance, and attitudes and behaviors.

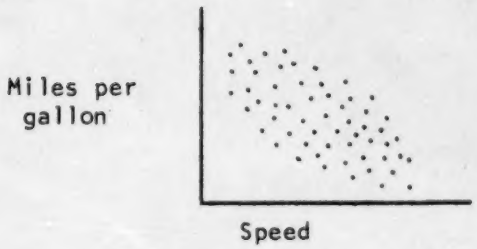
The following scatter diagrams illustrate some of the relationships that may be found. The first portrays a highly positive relationship, the second a strongly negative correlation, and the third a zero correlation.

Relationship Between Height and Weight



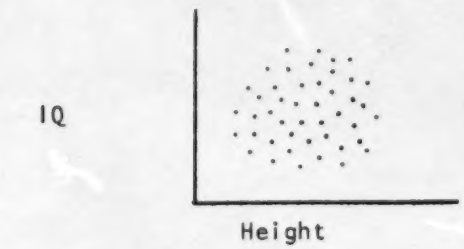
$r = .80$

Relationship Between Miles Per Gallon and Speed



$r = -.80$

Relationship Between IQ and Height



$r = 0$

As with the other statistics discussed, correlation coefficients may be interpreted in terms of statistical significance. The coefficients are checked against table values referencing the degrees of freedom for the subjects involved and the significance level selected. Where the obtained coefficient exceeds the table value, it is interpreted as representing a statistically significant relationship.

8. RANK ORDER CORRELATION. The product-moment correlation is used to calculate the relationship between two variables when direct measures of these variables are available. Sometimes it is not possible to use the actual scores, but only the ranks of each score: the highest, second highest, and so on. These ranks are known as ordinal data. The rank order correlation is a technique that can be used to test the relationship between two variables when only the ranks are known.

Suppose a researcher wanted to know whether pupil ranks in reading as determined by the teacher are similar to their ranks in arithmetic. The teacher's rankings of each student in reading and arithmetic provide the necessary data for calculation of the correlation. If each pupil ranks exactly the same in both reading and arithmetic, the rank order correlation will equal a perfect positive relationship. As with the product-moment correlation, rank order coefficients may range from +1.00 to -1.00.

The data for the rank order correlation can be organized as follows:

Student	Rank on Reading	Rank on Arithmetic
1	5	4
2	3	3
3	18	25
.	.	.
.	.	.
.	.	.
n	n	n

It should be mentioned that this coefficient can also be used when data obtained do not conform to the requirements of the product-moment correlation. As has been noted, the product-moment correlation assumes that the observed data are from interval or ratio variables.

When this requirement cannot be met, the rank-order correlation can be calculated as an approximation of the product-moment.

As with the product-moment correlation, it is possible to infer whether the value of the rank order correlation is statistically different from zero by comparing the obtained coefficient with the values found in the appropriate statistical table.

CONCLUDING NOTES

The data analysis techniques described in this guide are applicable to specific research hypotheses emerging from evaluation questions. Each evaluation implies a comparison, and each comparison indicates a statistical technique. Comparisons employed most frequently in educational evaluation are group performance vs. norms, group performance vs. previous performance, and group performance vs. another group's performance. Textbooks and other sources of information should be consulted before applying statistical tests in the various research and evaluation designs.

Data analysis assistance in each statistical procedure described in the guide is available through RBS Evaluation Services. Services include consultation with users on selection of appropriate techniques, processing of data for computer analysis and interpretation of results.

REFERENCES

The concepts and techniques discussed in the Analysis Service Guide are covered in detail in most elementary statistics textbooks. A few suggested references on the techniques and related areas are listed below.

A. Elementary Concepts

Guilford, J.P., and Fruchter, B. Fundamental statistics in psychology and education (5th Edition). New York, McGraw-Hill, 1973.

Runyon, R.P. and Haber, A. Fundamentals of behavioral statistics. Reading, Massachusetts: Addison-Wesley, 1967.

B. Analysis of Variance/Covariance

Edwards, A.L. Experimental design in psychological research (4th Edition). New York: Holt, 1972.

Winer, B.J. Statistical principle in experimental design (2nd Edition). New York: McGraw-Hill, 1971.

C. Multivariate Statistics

Kerlinger, F.N. and Pedhazur, E.J. Multiple regression in behavioral research. New York: Holt, Rinehart & Winston, 1973

Cooley, W.W. and Lohnes, P.R. Multivariate data analysis. New York: Wiley, 1971.

Morrison, D.F. Multivariate Statistical Methods. New York 1976.

D. Research Design

Campbell, D.T. and Stanley, J.C. Experimental and quasi-experimental design for research. Chicago: Rand McNally, 1966.

Bracht, G.H., and Glass, G.V. The external validity of experiments. American Educational Research Journal, 1968, 5, 437-474.

E. Evaluation Concepts

Worthen, B. and Sanders, J. Educational evaluation: theory and practice. Worthington, Ohio: Charles A. Jones, 1973.

Suchman, E.A. Evaluative research: principles and practice in public service and social action programs. New York: Russell Sage Foundation, 1967.

F. Educational Measurement

Cronbach, L.J. Essentials of psychological testing (3rd Edition). New York: Harper, 1970.

Thorndike, R.L. (Ed.) Educational measurement (2nd Edition). Washington, D.C.: American Council on Education, 1971.